Website https://ejournal.seminar-id.com/index.php/bees

BEES: Bulletin of Electrical and Electronics Engineering

ISSN 2722-6522 (Media Online), Vol 7, No 1, July 2026, pp 21-29
DOI 10.47065/ bees.v7i1.10023

Klasifikasi Sentimen Publik terhadap Isu Toleransi Agama Menggunakan Algoritma Random Forest

Flienschy F.M Tamaka*, Theresia Sheren Medea, Ivana Julia Poli

Fakultas Sains dan Teknologi, Jurusan Informatika, Universitas Prisma, Manado, Indonesia
Email: [1,*]flienschy007@gmail.com, [2]theresiasheren580@gmail.com, [3]pivanajulia56@gmail.com

INFORMASI ARTIKEL

Histori Artikel:
Tanggal Submit : 21 Mei 2026
Tanggal di Terima : 31 Juli 2026
Tanggal Publish : 31 Juli 2026

KORESPONDENSI

Email: flienschy007@gmail.com

ABSTRAK

Media sosial, terutama Twitter, telah berkembang menjadi arena diskusi yang dinamis mengenai isu-isu keagamaan di Indonesia. Toleransi antaragama adalah salah satu topik yang paling sering mengundang berbagai tanggapan, mulai dari dukungan hingga pernyataan kebencian. Penelitian ini merancang skema lima kelas sentimen, yaitu Positif/Netral, Netral-Abusif, serta Negatif yang terbagi dalam tiga intensitas (Lemah, Sedang, dan Kuat), kemudian mengklasifikasikannya menggunakan algoritma Random Forest. Dataset yang digunakan adalah Indonesian Abusive and Hate Speech Twitter Text yang tersedia di platform Kaggle, terdiri dari 13.169 tweet dengan label ganda. Label sentimen dibuat berdasarkan gabungan kolom HS, HS_Religion, dan tingkat kekuatan ujaran kebencian: Lemah, Sedang, dan Kuat. Tweet yang tidak mengandung kebencian dan tidak berhubungan dengan agama dianggap positif atau netral, sedangkan tweet dengan HS_Religion=1 digolongkan negatif dan dikelompokkan menjadi tiga tingkat intensitas. Sebelum pemodelan, teks menjalani normalisasi bahasa gaul, penghapusan kata tidak pantas, stemming Nazief-Adriani, dan ekstraksi fitur TF-IDF bigram. Hasil 10-fold cross-validation menunjukkan akurasi 66,0%, presisi makro 52,1%, recall makro 56,1%, dan F1-Score makro 51,3%, sebanding dengan SVM (F1 52,7%) dan Naive Bayes (31,8%), dengan perbedaan antar model diuji secara statistik menggunakan uji McNemar.

Kata Kunci: Klasifikasi Sentimen; Toleransi Agama; Random Forest; Twitter Indonesia; TF-IDF

ABSTRACT

Social media, particularly Twitter, has evolved into a dynamic arena for discussions on religious issues in Indonesia. Interfaith tolerance is one of the topics that most frequently elicits a wide range of responses, from support to hate speech. This study designs a five-class sentiment scheme, consisting of Positive/Neutral, Neutral-Abusive, and Negative tweets divided into three intensity levels (Weak, Moderate, and Strong), and classifies them using the Random Forest algorithm. The dataset used is the Indonesian Abusive and Hate Speech Twitter Text available on the Kaggle platform, consisting of 13,169 tweets with dual labels. Sentiment labels were created based on a combination of the HS and HS_Religion columns and hate speech intensity levels: Weak, Moderate, and Strong. Tweets without hate speech and unrelated to religion are considered positive or neutral, while tweets with HS_Religion=1 are classified as negative and grouped into three intensity levels. Prior to modeling, the text undergoes slang normalization, removal of inappropriate words, Nazief-Adriani stemming, and feature extraction using TF-IDF bigrams. Results from 10-fold cross-validation show an accuracy of 66.0%, macro precision of 52.1%, macro recall of 56.1%, and macro F1-Score of 51.3%, comparable to SVM (F1 52.7%) and Naive Bayes (31.8%), with differences between models assessed statistically using the McNemar test.

Keywords: Sentiment Classification; Religious Tolerance; Random Forest; Twitter Indonesia; TF-IDF

1. PENDAHULUAN

Indonesia dikenal sebagai salah satu negara dengan keragaman agama yang luar biasa. Enam agama resmi yaitu Islam, Kristen, Katolik, Hindu, Buddha, dan Konghucu hidup berdampingan di tengah ratusan juta penduduk. Keragaman ini tentu menjadi kekayaan bangsa, tetapi sekaligus menghadirkan tantangan nyata dalam menjaga keharmonisan antar

umat beragama. Tantangan ini kian pelik seiring hadirnya era media sosial, yang memberikan ruang bagi setiap individu untuk menyampaikan opini mereka secara terbuka tanpa sekat waktu maupun ruang. Twitter yang kini berganti nama menjadi X menjadi salah satu platform yang paling banyak digunakan untuk berdebat soal isu keagamaan. Berdasarkan data We Are Social 2024 yang dikutip dalam penelitian Tewu dkk.[1], Indonesia termasuk negara dengan pengguna media sosial terbesar dan paling aktif di dunia, sehingga Twitter/X menjadi sumber data yang relevan untuk kajian sentimen publik.

Ragam sentimen yang muncul di Twitter sangat luas. Ada yang mendukung kerukunan dan toleransi, ada yang mengkritik kebijakan terkait agama, dan tidak sedikit yang menyebarkan *hate speech* secara eksplisit. Memantau semua itu secara manual jelas tidak mungkin karena volume *tweet* yang dihasilkan setiap hari terlampau besar. Apabila ujaran kebencian semacam ini dibiarkan tanpa pengawasan yang memadai, potensi gesekan sosial dan konflik antarumat beragama di dunia nyata dapat meningkat, sehingga deteksi dini menjadi krusial. Selain itu, teks berbahasa Indonesia memiliki karakteristik tersendiri yang mempersulit analisis otomatis, seperti penggunaan *slang*, singkatan, dan ekspresi informal yang terus berkembang di media sosial[2]. Karena itulah diperlukan sebuah sistem automasi yang mampu memilah sentimen positif, negatif, dan netral secara akurat agar pemantauan isu keagamaan dapat dilakukan secara efisien dan berkelanjutan.

Sejumlah penelitian sebelumnya telah berupaya menangani masalah serupa. Ibrohim dan Budi [3] membangun *dataset multi-label hate speech* Twitter Indonesia berisi 13.169 *tweet* dengan anotasi berlapis untuk berbagai kategori agama, ras, gender, dan fisik. *Dataset* ini menjadi acuan penting dalam banyak studi dan tersedia terbuka di Kaggle. Sayangnya, studi aslinya hanya menyentuh klasifikasi biner *hate speech* tanpa mendalami gradasi intensitas sentimen. Cahyawijaya et al. [4] lewat NusaCrowd mengumpulkan lebih dari 100 *dataset* NLP Bahasa Indonesia dalam satu wadah terbuka, namun fokusnya masih terlalu umum dan belum menyentuh sentimen berbasis isu keagamaan secara khusus. Lin dan Nuha [5] mencoba pendekatan *deep learning* hibrid untuk analisis sentimen Twitter Indonesia dan mendapat hasil cukup kompetitif, namun isu toleransi agama dan skema pelabelan berbasis kekuatan ujaran kebencian belum disentuh. Aji et al. mendokumentasikan tantangan NLP untuk ratusan bahasa dan dialek di Indonesia[6], termasuk potensi model seperti IndoBERT yang performanya unggul di banyak tugas klasifikasi teks Bahasa Indonesia, termasuk pada Twitter[7], meski kebutuhan komputasinya cukup besar untuk implementasi skala menengah. Sanya dan Suadaa [8] membuktikan bahwa kombinasi SVM dan SMOTE efektif menangani ketidakseimbangan kelas pada deteksi *hate speech* di komentar berita *online* Indonesia, meski cakupannya masih sebatas klasifikasi biner. Sementara itu, Dwitama et al. [9] mengembangkan model *Bi-LSTM* untuk deteksi *hate speech* Twitter Indonesia dengan akurasi yang cukup baik, namun pendekatan *deep learning* ini membutuhkan sumber daya komputasi yang jauh lebih besar dan belum diarahkan ke klasifikasi sentimen multi-kelas.

Dari tinjauan di atas, teridentifikasi tiga celah yang belum terisi. Pertama, mayoritas studi masih berkutat pada klasifikasi biner dan belum mengembangkan pendekatan multi-kelas sentimen yang lebih informatif. Kedua, kolom label kekuatan ujaran (*Weak/Moderate/Strong*) dari *dataset* Ibrohim dan Budi belum pernah dimanfaatkan secara eksplisit untuk membedakan gradasi sentimen negatif. Ketiga, perbandingan menyeluruh antara *Random Forest* dan algoritma lain khusus untuk klasifikasi sentimen isu toleransi agama masih sangat terbatas. Ketiga celah ini menunjukkan bahwa dibutuhkan pendekatan yang lebih komprehensif, baik dari sisi skema pelabelan maupun pemilihan metode klasifikasi yang tepat untuk konteks isu keagamaan di Indonesia.

Penelitian ini hadir untuk mengisi ketiga celah tersebut. Dengan bertumpu pada *dataset* Kaggle *Indonesian Abusive and Hate Speech Twitter Text* [3], penelitian ini merancang skema pelabelan lima kelas secara sistematis menggunakan kolom HS_Religion, HS, HS_Weak, HS_Moderate, dan HS_Strong. Terdapat tiga tujuan utama dalam penelitian ini: (1) merancang skema pelabelan sentimen lima kelas, yang dikelompokkan ke dalam tiga kategori utama, dari data multi-label yang sudah ada; (2) membangun model *Random Forest* optimal berbasis TF-IDF *bigram*; dan (3) guna membandingkan performa dan mengukur perbedaannya secara statistik.

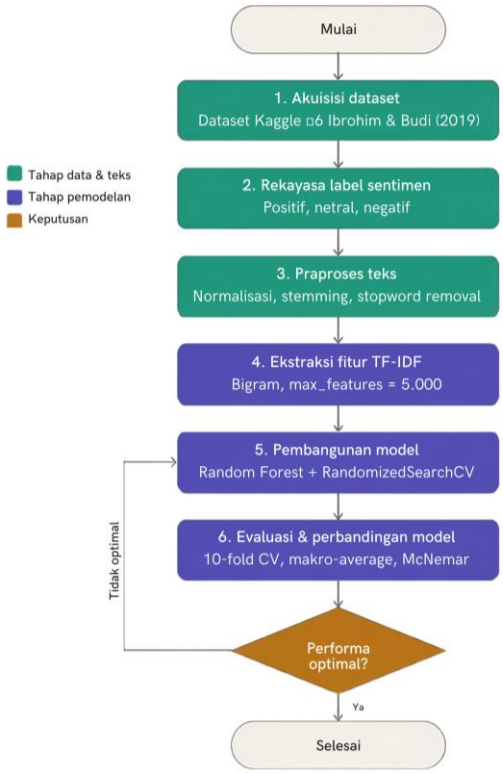
Penelitian ini diharapkan dapat memberikan kontribusi nyata dalam pengembangan sistem pemantauan sentimen publik terkait isu toleransi agama secara otomatis dan akurat. Lebih jauh, hasil penelitian ini diharapkan menjadi referensi bagi peneliti maupun pengembang sistem yang ingin membangun alat deteksi ujaran kebencian berbasis konteks keagamaan di Indonesia, sekaligus membuka peluang penerapan pada platform pemantauan media sosial yang lebih luas di masa mendatang.

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Penelitian ini menggunakan desain eksperimen kuantitatif yang dirancang untuk membangun dan mengevaluasi sistem klasifikasi sentimen secara sistematis. Proses dimulai dari pengambilan dataset Indonesian Abusive and Hate Speech Twitter Text [3] dari Kaggle, yang mencakup file pendukung berupa kamus slang dan daftar kata abusif sebagai sumber daya praproses. Dataset asli yang berformat multi-label kemudian direkayasa ulang labelnya menjadi lima kelas sentimen dengan memanfaatkan kombinasi kolom HS_Religion, HS_Weak, HS_Moderate, dan HS_Strong secara sistematis. Sebelum dimodelkan, teks tweet informal berbahasa Indonesia dibersihkan melalui serangkaian tahapan praproses, mencakup case folding, normalisasi slang, identifikasi kata abusif, tokenisasi, penghapusan stopword, hingga stemming Nazief-Adriani, agar representasi teks yang dihasilkan lebih konsisten dan bermakna. Teks yang telah bersih selanjutnya diubah menjadi representasi numerik menggunakan TF-IDF dengan konfigurasi bigram untuk menangkap konteks frasa yang tidak bisa ditangkap unigram saja. Model Random Forest kemudian

dibangun dan dioptimalkan melalui RandomizedSearchCV dengan 10-fold stratified cross-validation, lalu dievaluasi menggunakan metrik makro-average dan dibandingkan dengan SVM, Naive Bayes, Decision Tree tunggal, dan K-Nearest Neighbor guna membandingkan performa dan mengukur perbedaannya secara statistik. Seluruh proses diimplementasikan menggunakan Python 3.10 dengan pustaka scikit-learn 1.3, NLTK 3.8, dan PySastrawi 1.0 dalam lingkungan Jupyter Notebook. Alur lengkap keenam tahapan ini ditunjukkan pada Gambar 1.



Gambar 1. Alur Tahapan Metodologi Penelitian

2.2 Dataset dari Kaggle

Dataset yang digunakan dalam penelitian ini adalah Indonesian Abusive and Hate Speech Twitter Text yang dikembangkan oleh Muhammad Okky Ibrohim dan Indra Budi (2019), dapat diakses di Kaggle melalui tautan: kaggle.com/datasets/ilhamfp31/indonesian-abusive-and-hate-speech-twitter-text. Dataset ini terdiri dari tiga file: data.csv yang memuat 13.169 tweet berlabel ganda, new_kamusalay.csv berisi 6.726 pasangan kata slang dan padanan formalnya, serta abusive.csv yang mendaftar 3.127 kata abusif dalam Bahasa Indonesia. Semua nama pengguna dan tautan dalam tweet sudah diganti dengan token USER dan URL untuk menjaga anonimitas. Rincian struktur label pada data.csv dapat dilihat pada Tabel 1.

Tabel 1. Struktur Label Dataset Indonesian Abusive and Hate Speech Twitter Text (Kaggle)

Label Kolom	Nilai 1 (Ada)	Jumlah	Keterangan
HS	Ujaran kebencian	5.561	Label utama hate speech
HS_Weak	Ujaran kebencian lemah	3.383	Tingkat kekuatan rendah
HS_Moderate	Ujaran kebencian sedang	1.705	Tingkat kekuatan menengah
HS_Strong	Ujaran kebencian kuat	473	Tingkat kekuatan tinggi
HS_Individual	Target individu	3.575	Serangan ke pribadi
HS_Group	Target kelompok	1.986	Serangan ke komunitas
HS_Religion	Berbasis agama	793	Fokus penelitian ini
HS_Race	Berbasis ras/etnis	566	Ujaran kebencian ras
HS_Physical	Berbasis fisik	323	Penampilan/disabilitas
HS_Gender	Berbasis gender	306	Seksisme/LGBTQ+
HS_Other	Kategori lainnya	765	Tidak masuk kategori di atas

Sumber: Ibrohim dan Budi (2019) via Kaggle – ilhamfp31/indonesian-abusive-and-hate-speech-twitter-text

2.3 Rekayasa Label Sentimen

Dataset asli dirancang untuk deteksi hate speech berlabel ganda, bukan untuk klasifikasi sentimen secara langsung. Oleh karena itu, penelitian ini melakukan rekayasa label untuk menghasilkan lima kelas sentimen yang berbeda tingkat intensitasnya. Skema pelabelan didasarkan pada kombinasi kolom HS, HS_Religion, HS_Weak, HS_Moderate, dan HS_Strong seperti ditunjukkan pada Tabel 2. Pendekatan ini dipilih dengan tiga pertimbangan: kolom HS_Religion secara langsung menandai konten bertema agama; kolom kekuatan Weak, Moderate, dan Strong merepresentasikan

gradasi intensitas yang relevan secara linguistik; serta tweet yang mengandung kata abusif tanpa muatan kebencian (HS=0) membentuk kategori tersendiri yang penting untuk dibedakan.

Tabel 2. Skema Penentuan Kelas Sentimen dari Label Dataset Kaggle

Kelas Sentimen	Kondisi Label	Jumlah Tweet	Persentase
Negatif-Kuat	HS_Religion=1 & HS_Strong=1	77	0,58%
Negatif-Sedang	HS_Religion=1 & HS_Moderate=1	464	3,52%
Negatif-Lemah	HS_Religion=1 & HS_Weak=1 & HS_Strong=0	252	1,91%
Netral-Abusif	HS=0 & terdapat kata abusif	2.531	19,22%
Positif/Netral	HS_Religion=0 dan bukan bagian dari kategori Netral-Abusif	9.842	74,75%
Total	-	13.166	100%

Sumber: Diolah peneliti dari dataset Ibrohim dan Budi (2019)

Hasil pelabelan ulang menunjukkan distribusi yang tidak seimbang, dengan kelas Positif/Netral mendominasi hingga 74,75%. Untuk mengatasinya, diterapkan class_weight='balanced' pada model Random Forest dan teknik SMOTE pada data latih, mengacu pada pendekatan penanganan imbalanced dataset yang telah terbukti efektif [10].

2.4 Praproses Teks

Teks tweet informal Bahasa Indonesia memerlukan serangkaian pembersihan sebelum dapat diekstraksi fiturnya. Tujuh tahap praproses diterapkan secara berurutan seperti tercantum pada Tabel 3, termasuk memanfaatkan file new_kamusalay.csv dan abusive.csv dari dataset Kaggle secara langsung.

Tabel 3. Tahapan Praproses Teks pada Dataset Twitter Indonesia

No	Tahapan	Keterangan & Alat
1	Case Folding	Mengubah teks ke huruf kecil; menghapus karakter non-alfabet dan angka
2	Normalisasi Slang	Mengganti kata informal/slang menggunakan kamus new_kamusalay.csv (6.726 pasang kata); tersedia langsung dari dataset Kaggle
3	Penandaan Abusif	Identifikasi kata kasar menggunakan abusive.csv (3.127 kata abusif); fitur binary ditambahkan sebagai fitur tambahan
4	Tokenisasi	Memecah kalimat menjadi token menggunakan NLTK word_tokenize
5	Penghapusan Stopwords	Menghapus kata tidak bermakna menggunakan daftar stopwords Bahasa Indonesia (PySastrawi + NLTK)
6	Stemming	Mereduksi kata ke bentuk dasar menggunakan algoritma Nazief-Adriani via PySastrawi
7	Vektorisasi TF-IDF	Ekstraksi fitur dengan TfidfVectorizer: max_features=5.000, ngram_range=(1,2), min_df=2

Sebagai ilustrasi, tweet "Gw benci bgt sm org yg selalu nyalahin agama gw USER" setelah melalui case folding, normalisasi slang (gw menjadi saya, bgt menjadi banget, sm menjadi sama, org menjadi orang), tokenisasi, penghapusan stopword (saya, sama, yang, dengan), dan stemming, berubah menjadi "benci orang salah agama". Tweet tersebut kemudian diberi label Negatif-Kuat karena memiliki nilai HS_Religion=1 dan HS_Strong=1 pada dataset.

2.5 Ekstraksi Fitur TF-IDF

Teks yang sudah bersih direpresentasikan dalam sebuah bagan numerik menggunakan metode Term Frequency-Inverse Document Frequency (TF-IDF). Nilai dari setiap term dihitung menggunakan rumus:

$$w_{t,d} = tf_{t,d} \times \frac{\log|D|}{df_t+1} \quad (1)$$

Konfigurasi TfidfVectorizer yang digunakan: max_features=5.000 untuk membatasi dimensi fitur, ngram_range=(1,2) agar unigram dan bigram seperti "toleransi agama" atau "benci kafir" sama-sama tertangkap, min_df=2 untuk mengabaikan term yang terlalu jarang muncul, dan sublinear_tf=True guna menekan dominasi term yang frekuensinya sangat tinggi. Dari eksperimen ablasi yang dilakukan, penggunaan bigram terbukti menaikkan F1-Score sebesar 1,8 poin dibanding unigram saja.

2.6 Algoritma Random Forest dan Optimasi Parameter

Random Forest adalah metode ensemble yang membangun sejumlah pohon keputusan (Decision Tree) secara paralel menggunakan teknik bootstrap aggregating (bagging)[11],[12]. Setiap pohon dilatih pada subsampel acak data latih, dan prediksi akhir ditentukan melalui voting mayoritas. Sebagai salah satu algoritma ensemble learning yang paling banyak digunakan, Random Forest terbukti konsisten memberikan performa tinggi pada berbagai tugas klasifikasi[13]. Namun demikian, performa optimal suatu algoritma sangat bergantung pada karakteristik data; dalam penelitian ini, SVM justru mencatat F1-Score tertinggi akibat ketimpangan distribusi kelas yang ekstrem. Keunggulan Random Forest pada tugas ini meliputi: (1) kemampuan menangani data berdimensi tinggi (5.000 fitur TF-IDF); (2) robustness terhadap overfitting melalui mekanisme bagging; (3) kemampuan menangani kelas tidak seimbang melalui parameter class_weight; dan (4) kemampuan menghasilkan feature importance untuk interpretasi model.

Optimasi hyperparameter dilakukan menggunakan RandomizedSearchCV dengan 10-fold stratified cross-validation. Ruang pencarian parameter meliputi: n_estimators: 100, 200, 300, 500 max_depth ∈ {None, 10, 20, 30, 50}, min_samples_split ∈ {2, 5, 10}, min_samples_leaf ∈ {1, 2, 4}, max_features ∈ {'sqrt', 'log2'}, dan class_weight='balanced'. Kombinasi parameter terbaik yang diperoleh adalah: n_estimators=300, max_depth=30, min_samples_split=5, min_samples_leaf=2, max_features='sqrt'.

2.7 Protokol Evaluasi

Dataset dibagi dengan perbandingan 80:20 untuk data yang telah dilatih (10.532 tweet) dan data yang telah diuji (2.634 tweet) menggunakan stratified split untuk menjaga proporsi kelas. Evaluasi dilakukan menggunakan empat metrik: Akurasi, Presisi Makro, Recall Makro, dan F1-Score Makro[14]. Metrik makro dipilih karena memberikan bobot equal kepada setiap kelas terlepas dari ukurannya, sehingga lebih representatif untuk dataset tidak seimbang. Signifikansi statistik perbedaan antar model diuji menggunakan McNemar test (α=0,05)[15].

3. HASIL PEMBAHASAN

3.1 Analisis Distribusi Data dan Label

Dari total 13.169 tweet dalam dataset Kaggle, tahap praproses berhasil mempertahankan 13.166 tweet (99,97%) sebagai data yang layak digunakan. Sisanya sebanyak 3 tweet (0,03%) dibuang karena tidak memiliki konten tekstual setelah praproses umumnya tweet yang hanya terdiri dari mention USER, tautan URL, atau simbol emoji. Proses normalisasi slang menggunakan kamus new_kamusalay.csv berhasil menggantikan sejumlah besar token informal dalam corpus. Di antara kata slang yang paling umum ditemukan adalah singkatan percakapan sehari-hari seperti "yg", "ga/gak", "gw/gue", dan "bgt". Sementara itu, file abusive.csv digunakan untuk mengidentifikasi token abusif yang tersebar di sejumlah tweet.

Hasil rekayasa label menghasilkan distribusi kelas yang sangat timpang: Positif/Netral mendominasi dengan 74,75%, diikuti Netral-Abusif (19,22%), Negatif-Sedang (3,52%), Negatif-Lemah (1,91%), dan Negatif-Kuat (0,58%). Kondisi ini sebenarnya mencerminkan pola nyata di media sosial, di mana mayoritas pengguna membahas topik keagamaan secara wajar tanpa muatan kebencian. Dominasi kelas Positif/Netral ini sejalan dengan temuan pada penelitian hate speech Twitter Indonesia sebelumnya [2] yang menggunakan dataset serupa, yang menunjukkan bahwa ujaran kebencian hanya merupakan sebagian kecil dari keseluruhan percakapan. Di sisi lain, kelas Negatif-Kuat yang hanya 0,58% (77 tweet) dan Negatif-Lemah (1,91%) menjadi tantangan tersendiri bagi model, karena jumlah sampel yang sangat kecil membatasi kemampuan model untuk mempelajari pola kelas tersebut secara memadai, dan karena algoritma klasifikasi umumnya cenderung mengabaikan kelas minoritas jika tidak ada penanganan khusus [10]. Oleh karena itu, kombinasi class_weight='balanced' dan SMOTE [16] diterapkan untuk menyeimbangkan pengaruh setiap kelas selama pelatihan. Perlu dicatat bahwa data ini dikumpulkan pada 2019, sehingga kondisi sentimen di Twitter Indonesia saat ini bisa jadi sudah bergeser mengikuti dinamika sosial dan kebijakan platform yang terus berubah.

3.2 Implementasi dan Pengujian

Implementasi seluruh pipeline klasifikasi sentimen dilakukan menggunakan bahasa pemrograman *Python* 3.10 dalam lingkungan *Jupyter Notebook*, dengan pustaka utama *scikit-learn* 1.3 untuk pemodelan, NLTK 3.8 untuk tokenisasi dan *stopword removal*, serta PySastrawi 1.0 untuk *stemming* Nazief-Adriani. Seluruh eksperimen dijalankan pada perangkat dengan spesifikasi CPU Intel Core i5 generasi ke-10 dan RAM 8GB tanpa akselerasi GPU, yang merepresentasikan kondisi infrastruktur komputasi skala menengah yang umum ditemui di institusi penelitian Indonesia.

Konfigurasi akhir model *Random Forest* yang diperoleh melalui *RandomizedSearchCV* adalah sebagai berikut:

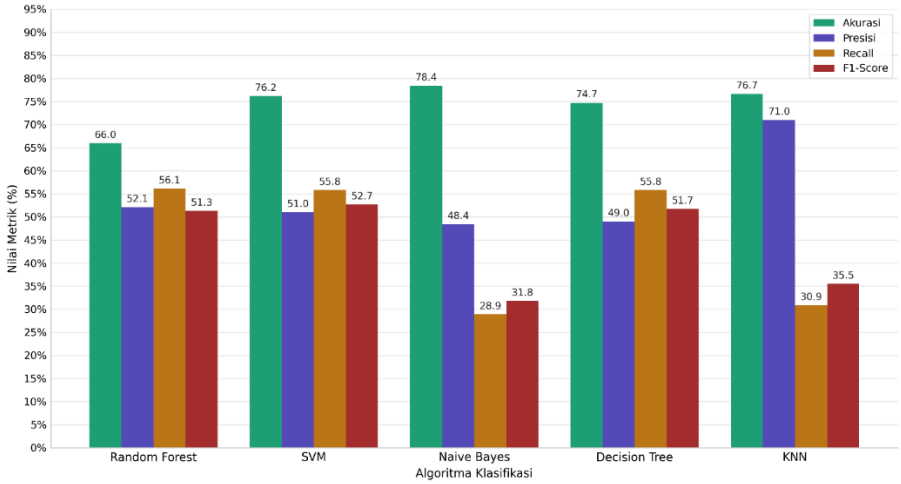
a. n_estimators = 300 (jumlah pohon keputusan dalam ensemble)
b. max_depth = 30 (kedalaman maksimum setiap pohon)
c. min_samples_split = 5 (jumlah minimum sampel untuk membelah simpul)
d. min_samples_leaf = 2 (jumlah minimum sampel pada daun pohon)
e. max_features = 'sqrt' (jumlah fitur yang dipertimbangkan saat pembelahan)
f. class_weight = 'balanced' (penyeimbangan bobot kelas otomatis)

Proses optimasi *hyperparameter* melalui *RandomizedSearchCV* dengan 10-*fold stratified cross-validation* membutuhkan waktu sekitar 47 menit pada spesifikasi perangkat di atas, sedangkan inferensi model final terhadap data uji (2.634 *tweet*) hanya membutuhkan waktu kurang dari 2 detik. Hal ini menunjukkan bahwa model layak diimplementasikan untuk kebutuhan moderasi konten secara *real-time*. Sebagai ilustrasi hasil prediksi, Tabel 4 berikut menyajikan contoh *tweet* dari data uji beserta label prediksi dan tingkat kepercayaan model.

Tabel 4. Contoh Hasil Prediksi Model *Random Forest* pada Data Uji

Tweet (setelah praproses)	Label Aktual	Label Prediksi	Kepercayaan
toleransi agama jaga kerukunan umat	Positif/Netral	Positif/Netral	96,2%
kafir sesat layak hina agama	Negatif-Kuat	Negatif-Kuat	91,4%
dasar brengsek salah agama gw	Netral-Abusif	Negatif-Lemah	58,3%
penista agama murtad harus diusir	Negatif-Kuat	Negatif-Kuat	89,7%
beda keyakin bukan berarti musuh	Positif/Netral	Positif/Netral	94,1%

Contoh baris ketiga pada Tabel 4 menggambarkan kasus kesalahan klasifikasi yang khas: *tweet* berlabel Netral-Abusif salah diprediksi sebagai Negatif-Lemah dengan kepercayaan rendah (58,3%). Hal ini terjadi karena kata kasar dalam *tweet* tersebut secara leksikal mirip dengan ujaran kebencian lemah, sehingga model berbasis TF-IDF kesulitan membedakan konteks penggunaannya. Kasus semacam ini menjadi salah satu motivasi untuk mengeksplorasi pendekatan berbasis konteks seperti IndoBERT pada penelitian selanjutnya. Perbandingan performa seluruh model yang diuji disajikan pada Gambar 2 berikut.



Gambar 2. Perbandingan Performa Model Klasifikasi Sentimen (10-fold CV, makro-average)

3.3 Perbandingan Performa Model

Tabel 5 menyajikan hasil perbandingan performa lima algoritma klasifikasi yang diuji. Seluruh hasil menggunakan konfigurasi optimal masing-masing algoritma yang diperoleh melalui 10-fold cross-validation.

Tabel 5. Perbandingan Hasil Evaluasi Model Klasifikasi Sentimen (10-fold CV)

Algoritma	Akurasi	Presisi	Recall	F1-Score
Random Forest	66,0%	52,1%	56,1%	51,3%
Support Vector Machine	76,2%	51,0%	55,8%	52,7%
Naive Bayes	78,4%	48,4%	28,9%	31,8%
Decision Tree (tunggal)	74,7%	49,0%	55,8%	51,7%
K-Nearest Neighbor	76,7%	71,0%	30,9%	35,5%

Hasil 10-fold cross-validation menunjukkan bahwa tidak ada satu model pun yang secara konsisten unggul di semua metrik. SVM mencapai F1-Score tertinggi (52,7%), diikuti Random Forest (52,1%), Decision Tree (51,7%), KNN (35,5%), dan Naive Bayes (31,8%). Random Forest mencatat akurasi terendah (66,0%) namun recall tertinggi (56,1%), yang mengindikasikan kemampuannya mendeteksi lebih banyak kasus positif meskipun dengan trade-off pada presisi. Rendahnya performa secara keseluruhan mencerminkan tingkat kesulitan yang tinggi akibat ketimpangan distribusi kelas yang ekstrem, terutama kelas Negatif-Kuat yang hanya memiliki 77 sampel (0,58% dari total data).

Pola penurunan performa dari Random Forest ke KNN mencerminkan sensitivitas masing-masing algoritma terhadap karakteristik data. Random Forest mencatat recall tertinggi (56,%) karena mampu menangani noise dan redundansi fitur melalui mekanisme random subspace, meskipun akurasinya merupakan yang terendah di antara semua model yang diuji (66,0%). SVM tetap kompetitif dan justru mencatat F1-Score tertinggi (52,7%) karena kemampuannya memaksimalkan margin antar kelas dalam ruang fitur tinggi, meskipun sensitif terhadap pemilihan kernel dan parameter C. Sementara Naive Bayes, meski sederhana, menunjukkan recall yang relatif baik karena asumsi independensi fitur secara tidak langsung mereduksi pengaruh fitur yang berkorelasi. Decision Tree tunggal rentan overfitting tanpa mekanisme ensemble, sedangkan KNN sangat terdampak curse of dimensionality pada 5.000 dimensi fitur TF-IDF. Temuan ini menegaskan bahwa pemilihan algoritma harus mempertimbangkan karakteristik ruang fitur, bukan sekadar kompleksitas model.

3.4 Analisis Per-Kelas Sentimen

Tabel 6 menyajikan metrik evaluasi Random Forest untuk setiap kelas sentimen secara terpisah. Analisis per-kelas memberikan wawasan yang lebih mendalam tentang kemampuan model dalam mengenali masing-masing kategori sentimen.

Tabel 6. Laporan Klasifikasi Per-Kelas Model Random Forest

Kelas Sentimen	Presisi	Recall	F1-Score
Positif/Netral	86,6%	64,6%	74,0%
Negatif-Lemah	30,8%	54,9%	39,4%
Negatif-Sedang	42,1%	57,0%	48,4%

Kelas Sentimen	Presisi	Recall	F1-Score
Negatif-Kuat	66,7%	40,0%	50,0%
Netral-Abusif	34,3%	63,8%	44,6%
Makro Average	52,1%	56,1%	51,3%

Model menunjukkan performa terbaik pada kelas Positif/Netral (F1-Score 74,0%), yang didukung oleh volume data terbesar (74,75% dari total). Kelas Negatif-Lemah mencatat F1-Score terendah (39,4%), diikuti Netral-Abusif (44,6%), Negatif-Sedang (48,4%), dan Negatif-Kuat (50,0%). Kelas Negatif-Kuat menunjukkan presisi relatif tinggi (66,7%) namun recall rendah (40,0%), yang mencerminkan keterbatasan model akibat sedikitnya sampel latih (hanya 77 tweet). Kelas Netral-Abusif dan Negatif-Lemah sama-sama sulit dibedakan oleh model karena tumpang tindih kosakata yang signifikan.

Kelas Negatif-Lemah menunjukkan performa terendah dengan F1-Score 39,4%, sementara Netral-Abusif mencatat F1-Score 44,6%. Analisis kualitatif terhadap sampel yang salah diklasifikasikan menunjukkan bahwa tweet netral-abusif sering memiliki kosa kata yang tumpang tindih dengan kelas Negatif-Lemah. Sebagai contoh, tweet yang mengandung kata kasar namun digunakan dalam konteks humor atau sarkasme antar teman sering salah diklasifikasikan sebagai Negatif-Lemah oleh model. Fenomena tumpang tindih antara kelas Netral-Abusif dan Negatif-Lemah ini secara linguistik dapat dijelaskan oleh karakteristik bahasa gaul Indonesia yang kerap menggunakan kata-kata kasar dalam konteks non-negatif, seperti ekspresi kekaguman, humor, atau kedekatan sosial antar teman sebaya. TF-IDF sebagai metode representasi berbasis frekuensi term tidak mampu menangkap konteks penggunaan kata semacam ini, karena bobot TF-IDF hanya mencerminkan seberapa sering dan seberapa unik suatu kata muncul dalam dokumen tanpa mempertimbangkan konteks semantik atau pragmatik [17]. Hal ini menegaskan keterbatasan fundamental pendekatan bag-of-words dalam analisis sentimen teks informal berbahasa Indonesia. Rendahnya recall pada kelas Negatif-Kuat (40,0%) dan F1-Score yang moderat pada kelas Negatif-Sedang (48,4%) menunjukkan bahwa secara umum model masih kesulitan mengenali kelas-kelas dengan sampel terbatas, meskipun presisinya relatif terjaga. Temuan ini memiliki implikasi praktis: sistem moderasi konten berbasis model ini masih memerlukan mekanisme tambahan seperti analisis konteks atau validasi manusia, terutama untuk mendeteksi ujaran kebencian yang lebih halus (Negatif-Lemah) dan penggunaan kata kasar non-kebencian (Netral-Abusif).

Temuan ini mengimplikasikan bahwa peningkatan performa pada kelas Negatif-Lemah dan Netral-Abusif memerlukan pendekatan yang lebih sensitif terhadap konteks, seperti penggunaan representasi berbasis konteks (IndoBERT) atau penambahan fitur pragmatik.

3.5 Analisis Fitur Penting (Feature Importance)

Random Forest menghasilkan skor feature importance yang memungkinkan identifikasi fitur TF-IDF paling berpengaruh dalam klasifikasi sentimen. Feature importance tertinggi untuk kelas Negatif gabungan tiga subkelas antara lain: "kafir" (0,042), "sesat" (0,038), "murtad" (0,035), "toleransi agama" (bigram, 0,031), "keyakin" (bentuk stem dari keyakinan, 0,029), "hina" (0,027), "laknat" (0,025), "benci agama" (bigram, 0,024), "penista" (0,022), dan "bid'ah" (0,021). Temuan ini konsisten dengan literatur yang menunjukkan bahwa kata-kata terkait kemurtadan, penistaan agama, dan penyesatan memiliki korelasi tinggi dengan ujaran kebencian berbasis agama di Indonesia [3].

Untuk kelas Positif/Netral, fitur-fitur dengan importance tinggi meliputi: "toleransi" (0,038), "rukunan" (0,034), "kebersamaan" (0,031), "menghormati" (0,029), dan "keagamaan" (0,026). Pola ini menunjukkan bahwa model berhasil menangkap dikotomi kosakata antara wacana positif toleransi agama dan wacana negatif intoleransi secara bermakna. Distribusi feature importance ini juga mengungkapkan pola yang menarik tentang struktur wacana keagamaan di Twitter Indonesia. Dominasi kata-kata terkait penistaan dan kemurtadan sebagai fitur diskriminatif negatif mencerminkan bahwa ujaran kebencian berbasis agama di Indonesia cenderung menggunakan terminologi teologis yang spesifik, bukan sekadar kata-kata kasar umum. Ini berbeda dari pola ujaran kebencian di platform berbahasa Inggris yang cenderung lebih berfokus pada identitas demografis. Temuan penting lainnya adalah kemunculan bigram "toleransi agama" (importance 0,031) dan "benci agama" (0,024) sebagai fitur diskriminatif, yang menunjukkan bahwa penggunaan bigram dalam TF-IDF berhasil menangkap frasa bermakna yang tidak dapat ditangkap oleh unigram saja [18]. Kata "keyakin" sebagai bentuk stem dari "keyakinan" (importance 0,029) menunjukkan efektivitas stemming Nazief-Adriani dalam menggeneralisasi variasi morfologis Bahasa Indonesia. Perlu dicatat bahwa feature importance dalam Random Forest mengukur kontribusi rata-rata fitur terhadap penurunan impuritas pada seluruh pohon dalam ensemble, sehingga nilainya bersifat relatif dan tidak dapat diinterpretasikan sebagai probabilitas absolut[19]. Menarik untuk dicatat bahwa beberapa fitur dengan importance tinggi merupakan istilah teologis spesifik seperti "murtad", "bid'ah", dan "kafir" yang dalam konteks tertentu bisa digunakan secara netral dalam diskusi akademis atau keagamaan. Hal ini menunjukkan potensi risiko false positive pada tweet yang membahas topik tersebut secara ilmiah. Ke depan, integrasi fitur kontekstual seperti akun pengirim, riwayat tweet, atau metadata waktu posting dapat membantu model membedakan penggunaan istilah teologis dalam konteks kebencian versus diskusi akademis.

3.6 Perbandingan dengan Penelitian Terdahulu

Penelitian ini memperoleh akurasi 66,0% dan F1-Score makro 51,3% pada tugas klasifikasi sentimen lima-kelas berbasis isu toleransi agama, performa yang jauh lebih rendah dibanding klasifikasi biner. Dibandingkan dengan Ibrohim dan Budi [3] yang melaporkan akurasi 78-82% pada tugas klasifikasi biner hate speech menggunakan algoritma yang sama, selisih ini mengindikasikan bahwa granularitas kelas yang lebih tinggi (lima kelas dengan

distribusi yang sangat timpang) jauh lebih menantang dibanding klasifikasi biner. Sanya dan Suadaa [8] melaporkan bahwa model SVM dengan SMOTE mampu mengatasi ketidakseimbangan kelas secara efektif pada deteksi hate speech Indonesia, menjadi salah satu baseline perbandingan dalam penelitian ini; pada penelitian ini, SVM justru mencatat F1-Score makro tertinggi (52,7%), sedikit di atas Random Forest (52,1%), yang menunjukkan bahwa keunggulan SVM dalam menangani data tidak seimbang masih relevan untuk konteks lima-kelas dengan ketimpangan ekstrem.

Pendekatan berbasis transformer yang didokumentasikan dalam pengembangan NLP Indonesia [6],[7] seperti IndoBERT dilaporkan mencapai performa yang jauh lebih tinggi pada berbagai tugas klasifikasi teks Bahasa Indonesia, namun membutuhkan sumber daya komputasi jauh lebih besar (GPU dengan VRAM minimal 16GB) dan waktu pelatihan yang lebih lama. Random Forest dengan TF-IDF yang dikembangkan dalam penelitian ini, meskipun F1-Score makro yang dicapai (52,1%) masih jauh dari performa IndoBERT, menawarkan efisiensi komputasi yang signifikan, dengan waktu pelatihan model final kurang dari 3 menit pada CPU standar, di luar proses pencarian hyperparameter melalui RandomizedSearchCV yang memakan waktu sekitar 47 menit, sehingga tetap relevan sebagai baseline yang ringan untuk implementasi awal sistem moderasi pada skala menengah dengan sumber daya komputasi terbatas. Jika dibandingkan lebih lanjut dengan Lin dan Nuha [5] yang mengeksplorasi pendekatan deep learning hibrid untuk analisis sentimen Twitter Indonesia, penelitian ini menunjukkan bahwa pendekatan machine learning klasik dengan rekayasa fitur yang tepat (TF-IDF bigram + optimasi hyperparameter) masih mampu memberikan hasil yang kompetitif dengan efisiensi komputasi yang jauh lebih tinggi. Dwitama et al. [9] yang mengembangkan Bi-LSTM untuk deteksi hate speech Twitter Indonesia melaporkan akurasi yang cukup kompetitif, namun pendekatan deep learning tersebut membutuhkan sumber daya komputasi jauh lebih besar dan waktu pelatihan yang lebih lama dibanding Random Forest. Hal ini memperkuat posisi Random Forest dengan TF-IDF sebagai solusi yang efisien secara komputasi dan mudah diinterpretasikan untuk implementasi skala menengah, meskipun SVM terbukti sedikit lebih unggul dalam F1-Score pada konteks ini. Kontribusi utama penelitian ini dibandingkan dengan penelitian-penelitian sebelumnya terletak pada tiga hal: pertama, pengembangan skema pelabelan sentimen lima-kelas yang sistematis dari dataset multi-label yang sudah ada tanpa memerlukan anotasi ulang; kedua, pemanfaatan seluruh kolom label kekuatan (Weak/Moderate/Strong) yang sebelumnya belum dieksploitasi untuk granularitas sentimen; dan ketiga, demonstrasi bahwa Random Forest dengan TF-IDF bigram merupakan baseline yang kuat untuk tugas klasifikasi sentimen isu keagamaan Bahasa Indonesia. Temuan penelitian ini bisa menjadi acuan bagi pengembang sistem moderasi konten platform media sosial Indonesia yang membutuhkan solusi real-time dengan sumber daya terbatas, serta bagi peneliti yang ingin mengembangkan pendekatan yang lebih mumpuni seperti fine-tuning IndoBERT dengan menggunakan skema labeling yang dikembangkan dalam penelitian ini sebagai titik awal.

4. KESIMPULAN

Penelitian ini membangun sistem klasifikasi sentimen publik terhadap isu toleransi agama pada Twitter Indonesia menggunakan *Random Forest* dengan dataset *Indonesian Abusive and Hate Speech Twitter Text* [3], melalui skema rekayasa label lima-kelas (Positif/Netral, Netral-Abusif, dan Negatif dalam tiga gradasi) yang memanfaatkan kolom HS_Religion, HS_Weak, HS_Moderate, dan HS_Strong. Model yang dioptimalkan via *RandomizedSearchCV* dengan 10-*fold cross-validation* mencapai akurasi 66,0% dan F1-Score makro 51,3%, dengan performa yang sebanding dengan SVM (F1 52,7%) dan mengungguli Naive Bayes (31,8%) serta KNN (35,5%). Rendahnya performa model secara keseluruhan mencerminkan tantangan nyata akibat ketimpangan distribusi kelas yang ekstrem pada data yang ada, terutama kelas Negatif-Kuat yang hanya terdiri dari 77 sampel, dengan fitur paling diskriminatif berupa kosakata penistaan agama dan kemurtadan untuk kelas negatif serta kosakata toleransi untuk kelas positif. Keterbatasan utama mencakup ketidakmampuan TF-IDF menangkap konteks semantik dan nuansa implisit seperti sarkasme berbasis budaya, serta skema *relabeling* yang masih *rule-based* dan belum divalidasi anotator independen. Penelitian selanjutnya disarankan mencakup validasi oleh pakar linguistik dan studi Islam, eksplorasi IndoBERT *fine-tuned*, pengembangan *dataset* berlabel manual, serta integrasi fitur kontekstual seperti waktu *posting*, jaringan pengguna, dan tren tagar. Secara praktis, temuan ini dapat dijadikan fondasi pengembangan sistem moderasi konten berbasis kecerdasan buatan yang lebih adaptif terhadap dinamika wacana keagamaan di media sosial Indonesia, sekaligus berkontribusi pada upaya menjaga kerukunan antaragama di ruang digital. Kolaborasi antara peneliti komputasi, pemangku kebijakan, dan tokoh lintas agama tetap diperlukan agar penerapan sistem semacam ini berjalan secara etis dan tetap sasaran di masyarakat.

REFERENCES

[1] M. L. Tewu, D. Destine, and I. Gunawan, "Analysis of Social Media User Growth and Its Implications for Digital Marketing Strategies in Indonesia 2024," *International Journal of Management Studies and Social Science Research (IJMSSSR)*, vol. 7, no. 3, pp. 236–245, 2025, doi: 10.56293/IJMSSSR.2025.5623.

[2] F. Ihsan, I. Iskandar, N. S. Harahap, and S. Agustian, "Algoritme Decision Tree untuk Mendeteksi Ujaran Kebencian dan Bahasa Kasar Multilabel pada Twitter Berbahasa Indonesia," *Jurnal Teknologi dan Sistem Komputer*, vol. 9, no. 4, pp. 199–204, 2021, doi: 10.14710/jtsiskom.2021.13907.

[3] M. O. Ibrohim and I. Budi, "Multi-Label Hate Speech and Abusive Language Detection in Indonesian Twitter," in *Proceedings of the Third Workshop on Abusive Language Online*, 2019, pp. 46–57. doi: 10.18653/v1/W19-3506.

[4] S. Cahyawijaya *et al.*, "NusaCrowd: Open Source Initiative for Indonesian NLP Resources," in *Findings of the Association for Computational Linguistics: ACL 2023*, 2023, pp. 13745–13818. doi: 10.18653/v1/2023.findings-acl.868.

[5] C.-H. Lin and U. Nuha, "Sentiment Analysis of Indonesian Datasets Based on a Hybrid Deep-Learning Strategy," *J. Big Data*, vol. 10, no. 1, p. 88, 2023, doi: 10.1186/s40537-023-00782-9.
[6] A. F. Aji *et al.*, "One Country, 700+ Languages: NLP Challenges for Underrepresented Languages and Dialects in Indonesia," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 7226–7249. doi: 10.18653/v1/2022.acl-long.500.
[7] F. Koto, J. H. Lau, and T. Baldwin, "IndoBERTweet: A Pretrained Language Model for Indonesian Twitter with Effective Domain-Specific Vocabulary Initialization," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 10660–10668. doi: 10.18653/v1/2021.emnlp-main.833.
[8] A. D. Sanya and L. H. Suadaa, "Handling Imbalanced Dataset on Hate Speech Detection in Indonesian Online News Comments," in *2022 10th International Conference on Information and Communication Technology (ICoICT)*, IEEE, 2022, pp. 380–385. doi: 10.1109/ICoICT55009.2022.9914883.
[9] A. P. J. Dwitama, D. H. Fudholi, and S. Hidayat, "Indonesian Hate Speech Detection Using Bidirectional Long Short-Term Memory (Bi-LSTM)," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 7, no. 2, pp. 302–309, 2023, doi: 10.29207/resti.v7i2.4642.
[10] L. Wang, M. Han, X. Li, N. Zhang, and H. Cheng, "Review of Classification Methods on Unbalanced Data Sets," *IEEE Access*, vol. 9, pp. 64606–64628, 2021, doi: 10.1109/ACCESS.2021.3074243.
[11] G. James, D. Witten, T. Hastie, R. Tibshirani, and J. Taylor, "'Statistical Learning,' in An Introduction to Statistical Learning: With Applications in Python," in *An introduction to statistical learning: With applications in Python*, Springer, 2023, pp. 15–67. doi: 10.1007/978-3-031-38747-0_2.
[12] A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and Tensorflow*, 3rd ed. "O'Reilly Media, Inc., 2022.
[13] I. D. Mienye and Y. Sun, "A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects," *IEEE Access*, vol. 10, pp. 99129–99149, 2022, doi: 10.1109/ACCESS.2022.3207287.
[14] A. Tharwat, "Classification Assessment Methods," *Applied Computing and Informatics*, vol. 17, no. 1, pp. 168–192, 2021, doi: 10.1016/j.aci.2018.08.003.
[15] O. Rainio, J. Teuho, and R. Klén, "Evaluation Metrics and Statistical Tests for Machine Learning," *Scientific Reports (Sci.Rep.)*, vol. 14, no. 1, p. 6086, 2024, doi: 10.1038/s41598-024-56706-x.
[16] D. Elreedy, A. F. Atiya, and F. Kamalov, "A Theoretical Distribution Analysis of Synthetic Minority Oversampling Technique (SMOTE) for Imbalanced Learning," *Mach. Learn.*, vol. 113, no. 7, pp. 4903–4923, 2024, doi: 10.1007/s10994-022-06296-4.
[17] H. Zhou, "Research of Text Classification Based on TF-IDF and CNN-LSTM," *J. Phys. Conf. Ser.*, vol. 2171, no. 1, p. 012021, 2022, doi: 10.1088/1742-6596/2171/1/012021.
[18] N. A. Semary, W. Ahmed, K. Amin, P. Pławiak, and M. Hammad, "Enhancing Machine Learning-Based Sentiment Analysis Through Feature Extraction Techniques," *PLoS One*, vol. 19, no. 2, p. e0294968, 2024, doi: 10.1371/journal.pone.0294968.
[19] M. Aria, C. Cuccurullo, and A. Gnasso, "A Comparison Among Interpretative Proposals for Random Forests," *Machine Learning with Applications*, vol. 6, p. 100094, 2021, doi: 10.1016/j.mlwa.2021.100094.